International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 5, Issue 2, Apr 2015, 31-40

TRANS STELLAR
• Journal Publications • Research Consultancy

GENERATING POSITIVE-NEGATIVE RULES USING FUZZY FP-GROWTH & NAÏVE BAYES

SHIPRA KHARE & VIVEK JAIN

SRCEM, Gwalior, Madhya Pradesh, India

ABSTRACT

Mining is an important application through which analysis of various data can be done easily. Although there are various techniques implemented for the analysis, Rule mining is an efficient technique where the analysis is done on the basis of rules generated. Since Apriori and FP-Growth rule mining are the efficient techniques implemented so far in which frequent items are generated and so is the association rules from it. Rules generated from frequent items sets are important for analysis but sometimes infrequent item sets are also used for the analysis. Positive and Negative association rules generated from these frequent and In-frequent item sets enables the analysis of various dataset application. Here in this paper a hybrid combinatorial method of Fuzzy FP-Growth and then classification using Naïve Bayes is implemented for the generation of positive and negative rules from frequent and In-frequent item sets.

KEYWORDS: FP-Growth, Apriori, Fuzzy, Naïve Bayes, Frequent Item Sets, Candidate Sets, KDD

INTRODUCTION

Data Mining can be defined as method of extraction of useful information from large amount of data. Data mining sometimes is also termed as Knowledge Discovery in Databases (KDD). The constant development of information technology in human life is generating huge amount of data which is generally stored in formats like records, documents, images etc. The data collected from different applications thus requires proper mechanism of extracting knowledge from large repositories for better decision making. Data Mining thereby aims at the discovery of useful information from large collections of data in various repositories from various applications.

Data mining can be explained as process of analyzing data from different and multiple perspectives and summarizing or concluding it into useful and required information which is then used to enhance profits, cuts costs etc. Due to the wide availability and preventability of large amounts of data and the coming need for forming such data into useful information and knowledge data mining is necessary. Multiple organizations collect huge amounts of such data which is generally stored on tertiary storage and are slowly migrated to database systems. Due to these reasons database systems have limited success and the current database systems are unable to provide necessary functionality for user information who wants to take the advantage from this data through various techniques like itemsite mining etc.

Itemset mining is an examining data mining technique used for discovering or identifying valuable correlations in between data. The process involved in performing itemset mining focuses on discovering frequent itemsets which are patterns whose observed frequency of occurrence in the source data is higher than the given threshold. Frequent itemsets have application area in multiple real-life contexts like market basket analysis, medical image processing, biological data analysis etc. But multiple traditional approaches also ignores the interest of each item or transaction within the analyzed

data. Therefore for allowing treatement of items or transactions differently with being based on their relevance in the frequent itemset mining process the concept of weighted itemset is being used. Frequent itemset mining directs to associations discovery and correlation relationship between the conditions and items in large data bases or data sets having huge amounts of data. Thus this discovery of interesting correlation relationship among the items helps in multiple areas like medical diagnosis, gene regulatory network etc [1].

Frequent itemsets mining is important part of data mining and variant of association analysis like association rule mining, sequential pattern mining etc. In frequent itemsets mining itemsets are produced from big data sets by applying association rule mining algorithms like Partition method, Apriori technique, Incremental, Border algorithm Pincer-Search etc. but these take larger computing time for computing all the frequent itemsets. Extraction of frequent itemsets acts as a basic step in multiple association analysis techniques. An itemset is termed as a frequent itemset if it presents or is contained in a large enough portion of the dataset. This frequent occurrence of item is generally expressed in the form of support count. Thus complicated techniques for hiding or reforming users private information are required during a data gathering process. And also these techniques should never submit the correctness of mining results [2].

FP Growth algorithm used for discovering interesting correlations takes less time while searching the each level of the tree. The FP growth method alters the problem of finding long frequent patterns for searching the shorter ones recursively and then concatenating the suffix. It apply the least frequent items as a suffix providing good selectivity thereby substantially reducing the search cost of the method [3].

FP-Growth algorithm permits frequent itemset discovery without the generation of candidate itemset. Two step. It builds a compact data structure termed as FP-tree using two passes over the data set and extracts frequent itemsets directly from FP tree. In FP-growth algorithm mining is done on constructed conditional frequent pattern (FP) tree. This tree is extended prefix tree structure and stores crucial and quantitative information about multiple frequent sets. FP-growth method convert the problem of searching long frequent patterns to search for shorter once recursively and then focussing on suffix.

The association rule mining as an important component of data mining. Discovering association rules is the most important feature of data mining. Association rule mining which is generally used in medicine, biology, business etc. has wide application areas. Association rules are generally if/then statements which are capable of uncovering relationships between apparently unrelated data in relational database or any other information repository. An association rule consists of two parts, an antecedent (if) and a consequent (then). An antecedent can be explained as an item found in the data and a consequent is the item that is found in combination with the antecedent. Association rules are created with the help of data analyses for frequent if/then patterns and using support and confidence criteria for identifying important relationships. Support is explained as an indication of how frequently the items appear in the database and the Confidence indicates the number of times the if/then statements have been found or are true [4].

In data mining, association rules are generally used to analyze and predict customer behavior. It is important technique in shopping basket data analysis, product clustering, catalog design and store layout. With the help of association rules programs capable of machine learning can be build which is a type of artificial intelligence (AI) seeking to build programs having ability of becoming more efficient without need of being explicitly programmed.

Association rule learning or mining is a method for discovering interesting relations in between variables or the items in large databases. It recognizes strong rules discovered in databases with the help of different measures of interestingness. Based on the concept of these strong rules association rules for discovery of regularities between the products in large scale transaction data recorded through point of sale (POS) systems in supermarkets is done. The information thus obtained can be used as the basis for decisions for various activities like for e.g., promotional pricing or product placements etc. [5] Market basket analysis association rules are deployed in many other application areas including Web usage mining, intrusion detection, Continuous production, bioinformatics etc. Comparing with sequence mining association rule learning does not generally consider the order of items either inside a transaction or across the transaction.

In association rule mining on a set of transactions rules are found that predicts the occurrence of an item based on the occurrence of other items in the transaction. ARM is used to find hidden relationship between items. With the help of user specified threshold i.e. minimum support mining of association rules is able to predict the complete set of frequent patterns determining the complete set of frequent patterns with the minimum support given [8] and the user can also specify lower minimum support for retrieval of more correlations in the items.

ARM in market basket analysis analyzes the itemset which thereby is analyzed after customer purchasing. It is similar to the analysis of purchasing behavior of customer. Association rules are also used telecommunication networks, market, risk management and inventory control etc. with the help of ARM past transaction data can easily be analyzed for discovering customer purchasing behaviors for improving the quality of business decision. The association rules generally points out or describes the associations between items in the large database of the customer transactions.

RELATED WORK

L. Cagliero [1] et.al. explained that correlations which are frequently hold in data are represented by frequent weighted itemsets and for minimization of a cost function rare data correlations discovery is beneficial over mining frequent correlations. They remarked the issue related with rare and weighted itemsets discovery termed as Infrequent Weighted Itemset (IWI) mining problem. They proposed measures for IWI mining and algorithms capable of performing IWI and Minimal IWI mining efficiently. They resolved the issue related with infrequent itemsets discovery with the help of weights for differentiating between relevant items and not within each transaction proposing FPGrowth like algorithms which are capable of accomplishing IWI and MIWI mining. The discovered patterns are validated with data from real life context to check for its usefulness [1].

S. Nathiarasan [6] et.al. explained that association rule mining finds correlations between data Items which depends upon frequency of occurrence. They remarked that infrequent itemset mining is a dissimilarity of frequent itemset mining and discovers uninteresting patterns which are the data items occurring rarely. for discovering frequent itemset mining the weight for each individual item in transaction independent manner is considered. They reviewed multiple algorithms with respect to frequent and infrequent itemset used in association rule mining. They remarked that weighted itemset mining generally examines information mining system and uncovers profitable connections inside the information and fulfills infrequent itemset mining for profitable discovery of rare datasets in the transactions. The process lying behind states that frequent item set fining is followed by infrequent weighted item sets discovery. They suggested that algorithm MIWI has less computation time and increases performance efficiency in large database computing weighted transaction. [6].

G. Kaur [7] et.al. remarked that with the help of data mining association rules can be generated in large number of itemsets. They explained that association rule mining determines interesting relations in large databases between the variables which is used in decision making process. They presented the comparative performance of Apriori and FP-Growth algorithms used in association rule mining based on execution time of the algorithm and number of scans for different number of instances in the algorithm. They presented that with the help of association rules interesting patterns in the database can be found in multiple data mining applications. Defining the association rule mining they proposed that with its help interesting correlations, frequent patterns or associations among sets of items in the transaction databases, relational databases or other information repositories can b e determined [7].

D. Gupta [8] et.al. explained about generation of large number of rules based on support and confidence with the help of association rule mining. They also presented that due to large database size it is time consuming to find all the association rules from that database and sometimes the users are only interested in the associations among some items in the database therefore mining association rules should maximize the occurrences of useful pattern. They surveyed about negative and positive association rules form the infrequent itemset. They suggested in determining confident positive rules having strong correlation because the algorithm sometimes discovers negative association rules with strong negative correlation between the antecedents and consequents. Thus their proposed algorithm is capable of formulating both positive and negative association rules efficiently [8].

Haoyuan Li [9] et.al. explained Frequent itemset mining (FIM) which is used to extract co-current items frequently. They remarked that huge dataset size increases memory usage and computational cost. They actually proposed parallelization of FP Growth algorithm on distributed machines. Their proposed PFP is capable of partitioning computations in a manner that each machine can execute independent group of mining tasks thereby eliminating computational dependencies and communication in between the machines. Their scheme is capable of linear speedup virtually on large dataset providing scalability and support to query recommendation for search engines. Their algorithm is based on novel data and on computation distribution scheme virtually eliminating communication between computers and expressing the algorithm through Map Reduce model and it mines top-k patterns related to each item and does not depends upon user specified value used to global minimal support threshold [9].

M. H. Dunham [10] et.al. explained that association rules are useful for marketing and retail communities defining the outline of association rule. Defining the data mining they explained that with it hidden information can be found using knowledge discovery process via clustering, classification, prediction, and link analysis. Providing that association rules determine relationships in set of items in a database and the relationships are not affected by inherent properties of data and is based on co-occurrence of the data. With the help of association rules failure in telecommunications networks can be predicted through analysis of events. They laid their focus on basket market analysis. They surveyed multiple existing algorithms for generation of association rules determining and proposing important features of the algorithms and the data structures used in the algorithms giving the performance of the algorithms. With the help of parallel algorithms finding large itemsets task can be parallelized [10].

R. Agrawal [11] et.al. explained association rule mining in large datasets on the basis of market analysis of customers and buyers. In a market there are large databases of customer’s transactions containing items purchased by a customer in a visit. Thus association rules can be generated for the items in the database. Their algorithm incorporated buyer management, novel estimation and pruning techniques. They presented the problem related with mining of a large

set of basket data type transactions for association rules among sets of items with some least specified confidence. Thus the step taken by them is able to enhance databases with functionalities to process queries [11].

PROPOSED METHODOLOGY

Here the proposed methodology is based on the combinatorial method of rules generation and classification. The proposed methodology works in the following phases:

- Take an input dataset (Para a).
- Apply FP-Growth Association rule mining on the following dataset (Para b).
- Generate Fuzzy rules from the rules generated from FP Tree technique (Para c)
- The rules are then classified using Naïve Bayes Algorithm to generate final decision tree (Para d).

Para a

Here two types of dataset are used real life dataset and synthetic dataset. Here we use ARFF version of the dataset means attribute relation file format. Since the base work is for infrequent item sets, here we take both frequent and infrequent and weighted and un-weighted datasets.

Para b

Given a transaction database DB and a minimum support threshold, the problem of finding the complete set of frequent patterns is called the frequent pattern mining problem.

Step 1: Build a compact data structure called the FP-tree

- Built using 2 passes over the data-set.

Step 2: Extracts frequent itemsets directly from the FP-tree

Pass 1

- Scan data and find support for each item.
- Discard infrequent items.
- Sort frequent items in decreasing order based on their support.
- Use this order when building the FP-Tree, so common prefixes can be shared.

Pass 2

Nodes correspond to items and have a counter

- FP-Growth reads 1 transaction at a time and maps it to a path
- Fixed order is used, so paths can overlap when transactions share items (when they have the same prfix).
    - In this case, counters are incremented
- Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines)

- The more paths that overlap, the higher the compression. FP-tree may fit in memory.

Frequent itemsets extracted from the FP-Tree.

Para c

After the generation of rules from FP tree, Fuzzy is applied over these rules to generate minimum rules.

Fuzzy Sets for Quantitative Attributes

It is composed of three steps:

Step 1: Transform the original database into positive integer

Step 2: For each attribute

- Cluster values of the attribute i^{th} into k medoids
- Classify the attribute i^{th} into k fuzzy sets
- Generate membership functions for each fuzzy set
- End for

Step 3: Transform the database based on fuzzy sets

Para d

Generate Decision Tree from a set of rules using Naïve Bayes classifier.

RESULT ANALYSIS

The proposed methodology implemented here is tested on various popular datasets including:

- **Compendium Dataset:** The dataset is specially designed to investigate the various health services and research on medical education.
- **StemCells Dataset:** The dataset is based on Pluripotent cells which generates cell in the human body such as HESC.
- **Yeast Dataset:** The dataset contains the various predicted attributes of the protein localization. It contains 1484 number of instance values with 8 attributes.

The table shown below is the comparison of total frequent item sets generates on Step Cells Dataset. The Result is compared on Stem Cells Dataset for various support and confidence values.

Table 1: Comparison of Frequent Item Sets Generated

StemCell Dataset		Base Work	Proposed Work
Support	**Confidence**		
0.23	0.1	5	70
0.12	0.12	12	102
0.213	0.164	7	71
0.67	0.23	0	21
0.1	0.23	14	116

The figure shown below is the comparison of total frequent item sets generates on Step Cells Dataset. The Result is compared on Stem Cells Dataset for various support and confidence values.

The existing and proposed work is compared on various values of support and confidence. The rules generated for the existing and proposed work for support and confidence.

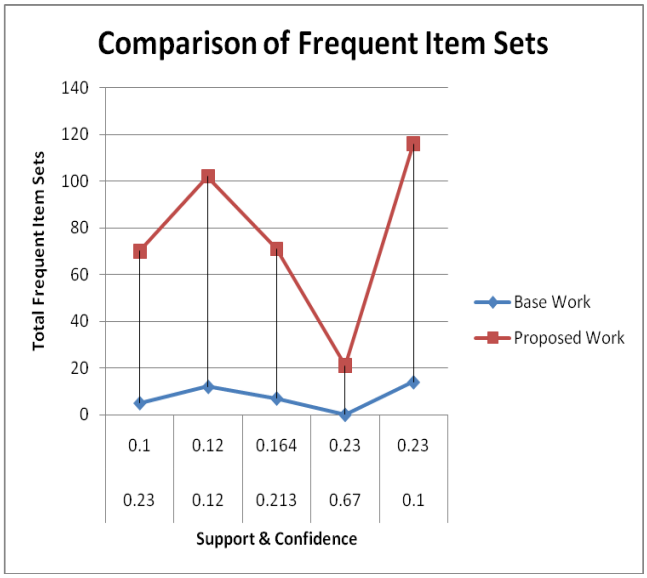

Figure 1: Comparison of Total Frequent Item Sets

The figure shown below is the comparison of total In-frequent item sets generates on Step Cells Dataset. The Result is compared on Stem Cells Dataset for various support and confidence values.

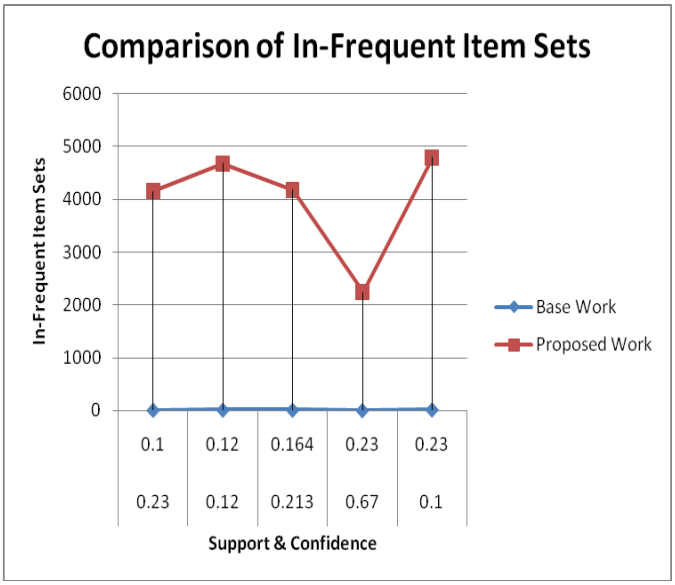

Figure 2: Comparison of Total in-Frequent Item Sets

The figure shown below is the comparison of total Positive rules generates on Step Cells Dataset. The Result is compared on Stem Cells Dataset for various support and confidence values.

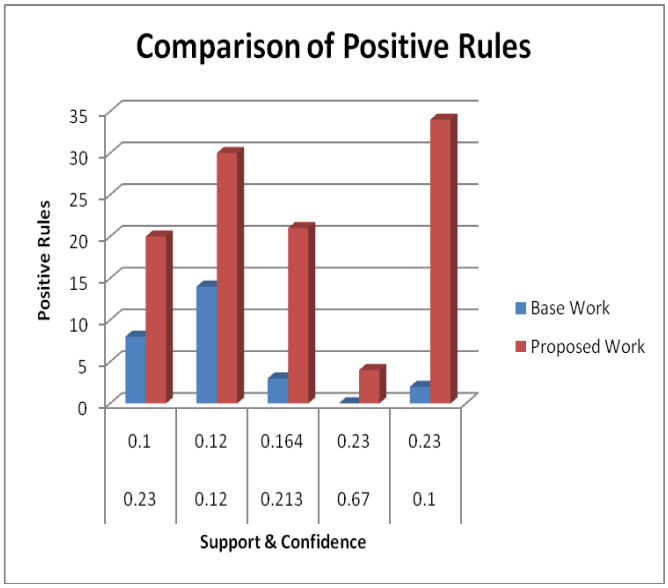

Figure 3: Comparison of Total Positive Rules Generated

The figure shown below is the comparison of total Negative rules generates on Step Cells Dataset. The Result is compared on Stem Cells Dataset for various support and confidence values.

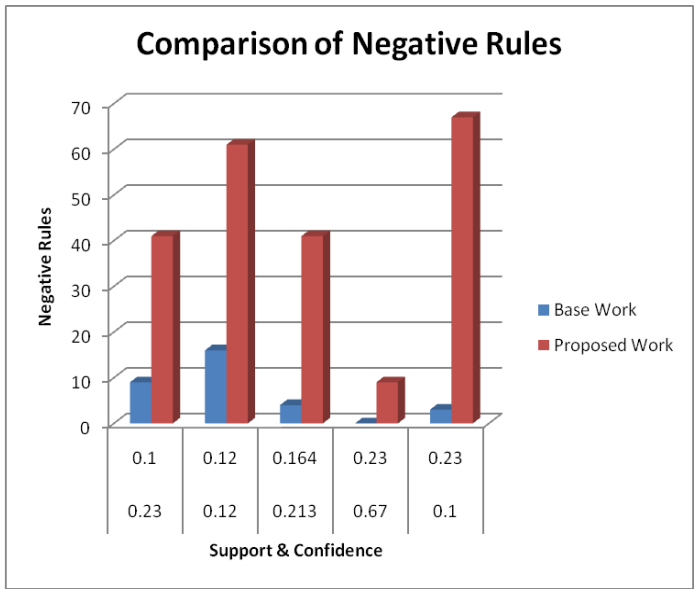

Figure 4: Comparison of Total Negative Rules Generated

The table shown below is the total execution time taken to generate rules and sets from the dataset.

Table 2: Comparison of Execution Time

StemCell Dataset		Base Work	Proposed Work
Support	Confidence		
0.23	0.1	0.4212	0.8268
0.12	0.12	0.1265	1.0452
0.213	0.164	0.3744	0.5148
0.67	0.23	0.2496	0.078
0.1	0.23	0.3658	1.17

CONCLUSIONS

The proposed methodology implemented here for the analysis of various datasets using Fuzzy FP-Growth and Naïve Bayes classifier is an efficient technique which generated positive and negative rules in a more classified manner, so that the analysis can be done easily and quickly. The experimental results are performed on various datasets and results are compared with the existing Rule generation techniques. The proposed methodology implemented has better performance as compared to the existing technique.

REFERENCES

1. Luca Cagliero and Paolo Garza "Infrequent Weighted Itemset Mining using Frequent Pattern Growth", IEEE 2013
2. James Cheng, Yiping Ke and Wilfred Ng "A Survey on Algorithms for Mining Frequent Itemsets over Data Streams", 2006.
3. Sujatha Dandu, B. L. Deekshatulu and Priti Chandra "Improved Algorithm for Frequent Item sets Mining Based on Apriori and FP –Tree", Global Journal of Computer Science and Technology Software & Data Engineering, 2013
4. Arvind Jaiswal, Gaurav Dubey "Identifying Best Association Rules and Their Optimization Using Genetic Algorithm", International Journal of Emerging Science and Engineering (IJESE), 2013
5. Zutao Zhu, Guan Wang, and Wenliang Du "Deriving Private Information from Association Rule Mining Results", 2007
6. Sakthi Nathiarasan, Kalaiyarasi and Manikandan "Literature Review on Infrequent Itemset Mining Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, 2014
7. Gagandeep Kaur and Shruti Aggarwal "Performance Analysis of Association Rule Mining Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering, 2013
8. Diti Gupta and Abhishek Singh Chauhan "Mining Association Rules from Infrequent Itemsets: A Survey", International Journal of Innovative Research in Science Engineering and Technology, 2013
9. Haoyuan Li, Yi Wang, Dong Zhang, Ming Zhang and Edward Chang "PFP: Parallel FP-Growth for Query Recommendation", ACM, 2007.
10. Margaret H. Dunham, Yongqiao Xiao, Le Gruenwald and Zahid Hossain "A Survey of Association Rules", 1999
11. Rakesh Agrawal, Tomasz Imielinski and Arun Swami "Mining Association Rules between Sets of Items in Large Databases:, 1993